DOCUMENT RESUME

ED 395 947 TM 025 014

AUTHOR Dorans, Neil J.; Lawrence, Ida M.
TITLE Checking the Equivalence of Nearly Identical Test Editions.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-88-6
PUB DATE Feb 88
NOTE 19p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Equated Scores; *Error of Measurement; Test Construction; *Test Format; Test Items
IDENTIFIERS *Equivalence Formation; Graphic Representation; *Scholastic Aptitude Test; Standard Deviation

ABSTRACT

A procedure for checking the score equivalence of nearly identical editions of a test is described. The procedure employs the standard error of equating (SEE) and utilizes graphical representation of score conversion deviation from the identity function in standard error units. Two illustrations of the procedure involving Scholastic Aptitude Test (SAT) data are presented. Results from this kind of analysis could have implications for equating practices and test assembly. If the two versions of the operational SAT test had differed by more than two SEEs, it might have been necessary to compute separate equating functions for each order and to consider taking an average of the two functions or using separate functions for each order depending on the practical consequences of the option. If the equating function for one or more of the item orders had been shown to differ by more than two SEEs, then test assemblers would have been alerted to the fact that item order affects test properties and that item order needs to be taken into account when assembling tests. (Contains two figures, two tables, and four references.) (Author/SLD)

RESEARCH REPORT

CHECKING THE EQUIVALENCE OF NEARLY IDENTICAL TEST EDITIONS

Neil J. Dorans

Ida M. Lawrence

Educational Testing Service
Princeton, New Jersey
February 1988

CHECKING THE EQUIVALENCE OF NEARLY IDENTICAL TEST EDITIONS

Neil J. Dorans and Ida M. Lawrence

College Board Statistical Analysis

Educational Testing Service

The authors wish to thank William Angoff, D. R. Divgi, Gene Johnson, Gary Marco, and David Wright for reviewing earlier drafts of this report.

Abstract

A procedure for checking the score equivalence of nearly identical editions of a test is described. The procedure employs the standard error of equating and utilizes graphical representation of score conversion deviation from the identity function in standard error units. Two illustrations of the procedure involving SAT data are presented.

CHECKING THE EQUIVALENCE OF NEARLY IDENTICAL TEST EDITIONS

Often a testing program may create two or more editions of a test, such as the Scholastic Aptitude Test (SAT), that are comprised of the same test items given in different orders. For example, one version of the test might contain one section ordering, while another version of the same test might contain a different section ordering. The literature on item rearrangement and section ordering contains examples in which the order of sections or the arrangement of items has an effect on test scores (Leary and Dorans, 1985). Hence, it is sound psychometric practice to check for potential item arrangement or section order effects before concluding that two nearly identical versions of a test are interchangeable. Score equating models can be used for this check on equivalence. In this paper, a procedure for assessing equivalence is described and two illustrations employing Scholastic Aptitude Test (SAT) data are given.

Equating Different Versions of the Same Test

Several times a year two versions of the SAT are administered to what are presumed to be statistically equivalent groups of examinees that are obtained via a process known as spiraling. Spiraling refers to a data collection design in which every other examinee receives one version of the test. The two SAT versions under consideration here differ only with respect to the particular order in which the test is

administered. For example, in one order, the 40-item verbal section might precede the 45-item verbal section, while in the other order, the 40-item verbal section comes after the 45-item verbal section. Two versions of the test are used for test security reasons. On the face of it, this transposition of sections should not have any effect. In practice, it is wise to ascertain whether equivalence does obtain.

The procedure employed for ascertaining statistical equivalence of these versions is to equate one version to the other version and to check whether or not the identity function falls within a reasonable confidence interval placed around that equating. Since both versions of the test contain identical items given in different orders, it is customary to presume that the equating function is linear. Let X equal the test version designated to have the original order, and let Y designate the test version with the secondary order. The goal is to equate Y to X in some sample of data in which approximately half the examinees are given X while the remainder are given Y. It is assumed that the samples are representative of the population of interest and that score distributions within the samples are similar in shape. In the sample, the linear equating function is given by

$$(1) \quad X^*(y) = \bar{X} + s_x (y - \bar{Y}) / s_y,$$

where $\bar{X}$, s_x and $\bar{Y}$, s_y are the sample means and standard deviations of X and Y, respectively and $X^*(y)$ is the equating function for equating Y to X. Braun and Holland (1982, pp. 33 - 34) present

several expressions for the standard error of equating, including the special case under which both X and Y have normal distributions. Under that condition, the expression for the standard error of equating Y to X via (1) is

$$(2) \quad SEE = \{ (s^2_x / n_h) (2 + Z^2(y)) \}^{.5}$$

where $n_h = [.5 (n_X^{-1} + n_Y^{-1})]^{-1}$ is the harmonic mean of n_X and n_Y, and $Z(y) = (y - \bar{Y}) / s_y$. Operationally, this standard error can be used to place a confidence band around the equating function depicted in (1). For SAT equatings, the boundaries of the confidence band are defined by plus or minus two SEEs at each raw score point. The test for equivalence then reduces to determining whether or not the identity function falls within the region defined by the confidence band.

To assess equivalence it is convenient to compute the difference between the equating function and the identity transformation, and to divide this difference by the standard error of equating. If the resultant ratio falls within a bandwidth of plus or minus two, then the equating function is deemed to be within sampling error of the identity function. Plotting this ratio against raw score enables one to ascertain swiftly where, if anywhere, the equating function departs appreciably from the identity function. Several illustrations of the procedure for assessing equivalence of SAT editions and sections follow.

Illustrations With SAT Data

Different Section Orders

Two new forms of SAT-Verbal (Form 1 and Form 2), spiraled at a single national administration, actually contained the same operational sections, but given in different orders. To determine whether the two forms could be treated as interchangeable for the purposes of score reporting, a linear equating function was used to equate raw scores on Form 2 (n = 43,845) to raw scores on Form 1 (n = 48,639). As part of the equating process, standard errors of equating (SEEs) for random groups were estimated. The results of this analysis are presented in Table 1. The first two columns of this table contain raw scores on Form 2 (in five point intervals) and the respective equivalent score on Form 1. The third column contains differences between the two conversions. The SEEs for the various score points are reported in the fourth column, while the fifth column contains ratios of differences in conversions to their SEEs. The ratio indicates the size of the difference in standard error units.

Table 1 shows only slight differences between the equating function and the identity function, differences ranging from .04 to .31 raw score points. These differences follow a monotonically decreasing trend, with the largest difference occurring at the top of the range and the smallest difference occurring at the bottom of the range. Since these differences are less than two SEEs throughout the score range, the two orders of

Table 1

Test of Equivalence for the Same SAT-V Form
Given in Two Different Section Orders

FORM 2 RAW SCORE	FORM 1 EQUIVALENT	DIFF	SEE	RATIO
85	84.695	.305	.26304	1.16
80	79.708	.292	.24212	1.21
75	74.720	.280	.22165	1.26
70	69.733	.267	.20177	1.32
65	64.745	.255	.18269	1.39
60	59.758	.242	.16468	1.47
55	54.770	.230	.14812	1.55
50	49.783	.217	.13357	1.63
45	44.795	.205	.12173	1.68
40	39.808	.192	.11347	1.69
35	34.820	.180	.10960	1.64
30	29.833	.167	.11058	1.51
25	24.845	.155	.11628	1.33
20	19.858	.142	.12607	1.13
15	14.870	.130	.13908	.93
10	9.883	.117	.15451	.76
5	4.895	.105	.17170	.61
0	- .092	.092	.19018	.48
- 5	- 5.080	.080	.20961	.38
-10	-10.067	.067	.22974	.29
-15	-15.055	.055	.25041	.22
-20	-20.042	.042	.27149	.16

SAT-Verbal can be considered interchangeable for the purpose of score reporting. Further perspective on the magnitude of these differences is obtained by realizing that the SEE is always smaller than the standard error of measurement for SAT-Verbal, which for these forms is 4.57.

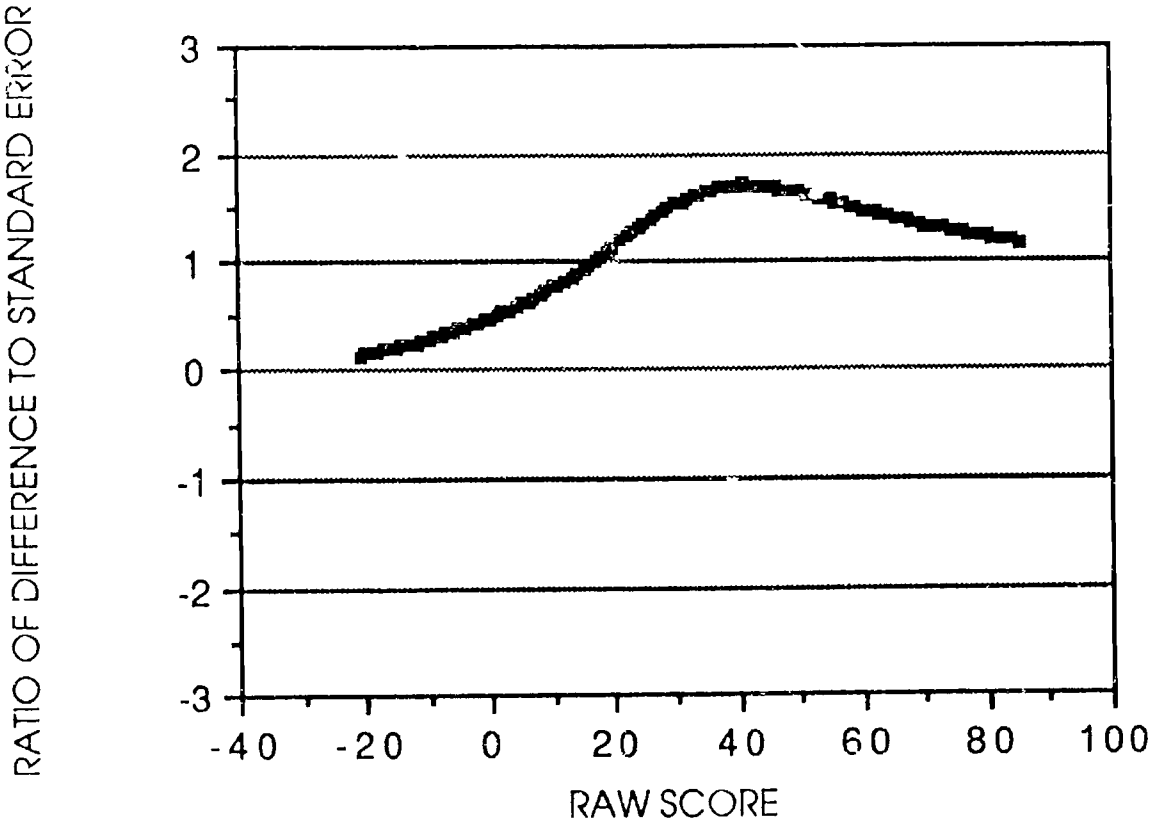
Note that the ratio displayed in the last column of Table 1 varies, in a nonmonotonic fashion, as a function of score level. This relationship is plotted in Figure 1. Although the largest raw score differences occur at the top of the score range, Figure 1 indicates that the differences, in standard error units, are largest in the middle of the range, between raw scores of 30 and 50. This is because the SEE is smallest for mid-range scores, where equating is more precise due to the preponderance of examinees scoring near the mean on either test.

Different Item Arrangements

As part of a study on the effects of item rearrangement (Harvey, 1987), five versions of the 35-item Mathematics 2 section the SAT-Mathematical were spiraled in the variable section for one national administration of the SAT. These sections differed only in the order of the Quantitative Comparison item type, which is located in positions 8 through 27 of the section. Five different item orders, for subforms A, B, C, D and E, respectively, were given, such that groups of approximately 8,000 examinees each took one of the five subforms.

FIGURE 1

TEST OF EQUIVALENCE OF DIFFERENT SECTION ORDERS



Subform A had the traditional item arrangement, while different rearrangements were employed for the other four subforms. The assumption of equivalent (for equating purposes) item orders was tested by equating raw scores on subforms B, C, D, and E to raw scores on subform A. As part of the equating, SEEs were estimated.

Results of these analyses are reported in Table 2, where each panel contains information pertaining to the different equatings. Based on the ratios of differences in conversions to SEE, it appears that the item orders in subforms B and D are most similar to the original item arrangement in subform A. Subform E differs somewhat from the original subform, and subform C is the most dissimilar. However, all of the differences are less that two SEEs, and much smaller than the the usual raw score standard error of measurement for the 35-item section of SAT-Mathemetical, which is typically 3.0.

The ratios as a function of score level for the four equatings are plotted in Figure 2, which provides an opportunity for comparing results across the four equatings. Note that the ratios for subforms B and D are near zero, indicating minimal differences between their respective conversions and the conversion for subform A. While the overall effect of rearranging items was minimal, the data for subforms C and E are further away from zero, suggesting that item rearrangement had a more pronounced effect for these subforms.

Table 2

Test of Equivalence for the Same SAT-M Sections With Different Item Orders:
Subforms B, C, D, and E Equated to Subform A

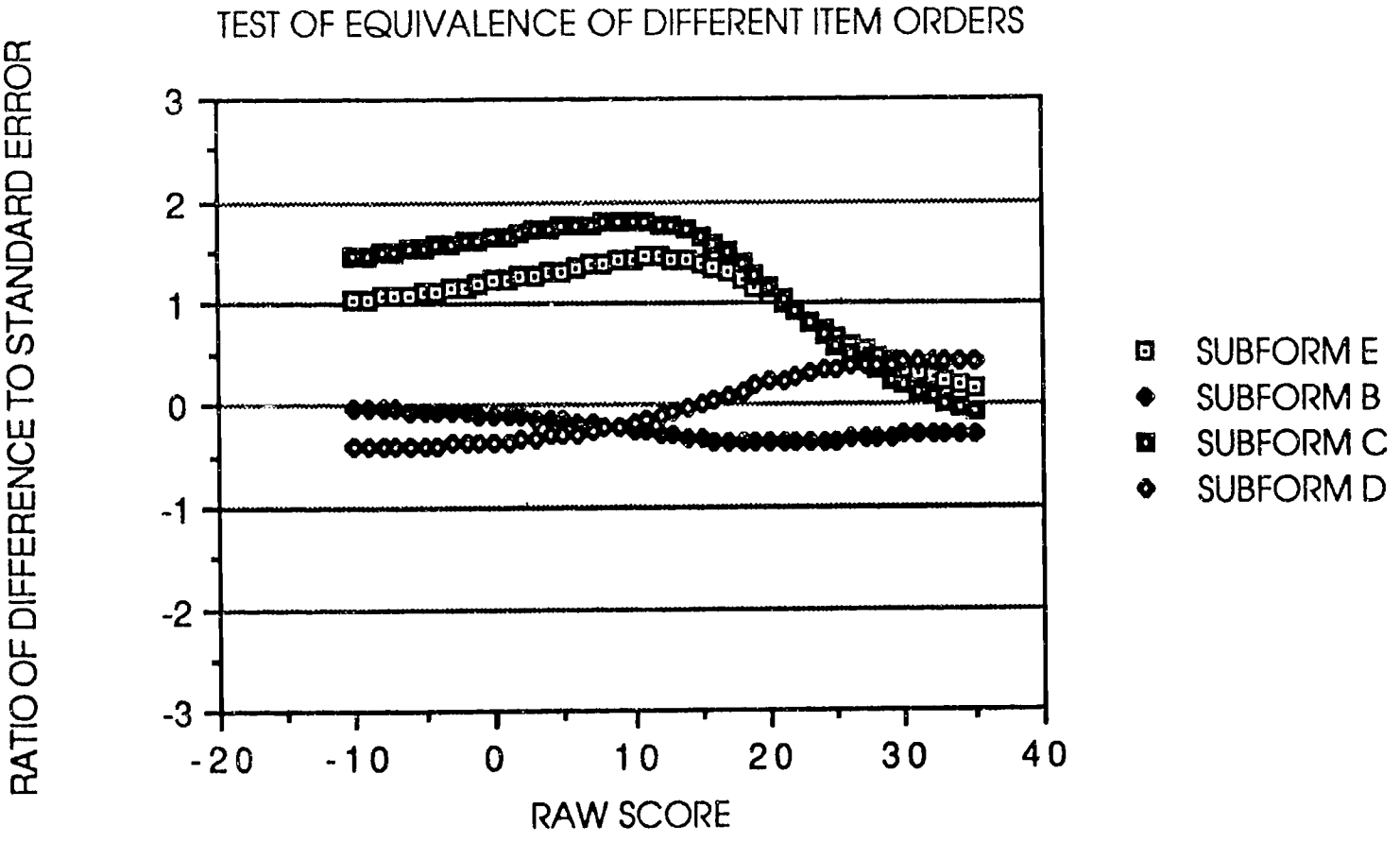
SUBFORM B RAW SCORE	SUBFORM A EQUIVALENT	DIFF	SEE	RATIO
35	35.064	-.064	.23568	-.27
30	30.058	-.058	.19040	-.30
25	25.051	-.051	.15138	-.34
20	20.045	-.045	.12463	-.36
15	15.038	-.038	.11876	-.32
10	10.032	-.032	.13649	-.23
5	5.025	-.025	.17061	-.15
0	0.019	-.019	.21341	-.09
-5	-4.988	-.012	.26064	-.05
-10	-9.994	-.006	.31028	-.02

SUBFORM C RAW SCORE	SUBFORM A EQUIVALENT	DIFF	SEE	RATIO
35	35.016	-.016	.23615	-.07
30	29.963	.037	.19034	.19
25	24.910	.090	.15101	.60
20	19.857	.143	.12447	1.15
15	14.804	.196	.11956	1.64
10	9.751	.249	.13860	1.80
5	4.698	.302	.17389	1.74
0	- .355	.355	.21767	1.63
- 5	- 5.408	.408	.26577	1.54
-10	-10.461	.461	.31623	1.46

SUBFORM D RAW SCORE	SUBFORM A EQUIVALENT	DIFF	SEE	RATIO
35	34.893	.107	.23713	.45
30	29.919	.081	.19169	.42
25	24.945	.055	.15263	.36
20	19.971	.029	.12600	.23
15	14.997	.003	.12037	.02
10	10.023	-.023	.13832	-.17
5	5.049	-.049	.17265	-.28
0	0.075	-.075	.21568	-.35
- 5	-4.899	-.101	.26317	-.38
-10	-9.873	-.127	.31310	-.41

SUBFORM E RAW SCORE	SUBFORM A EQUIVALENT	DIFF	SEE	RATIO
35	34.962	.038	.24220	.16
30	29.929	.071	.19534	.36
25	24.896	.104	.15512	.67
20	19.863	.137	.12795	1.07
15	14.830	.170	.12263	1.38
10	9.797	.203	.14219	1.43
5	4.764	.236	.17818	1.32
0	- 0.269	.269	.22291	1.21
- 5	- 5.302	.302	.27209	1.11
-10	-10.335	.335	.32371	1.03

FIGURE 2


TEST OF EQUIVALENCE OF DIFFERENT ITEM ORDERS
RATIO OF DIFFERENCE TO STANDARD ERROR
3
2
1
0
-1
-2
-3
-20
-10
0
10
20
30
40
RAW SCORE
SUBFORM E
SUBFORM B
SUBFORM C
SUBFORM D

Discussion

A procedure for checking the score equivalence of nearly identical editions of a test has been described. The procedure employs the standard error of equating and utilizes graphical representation of score conversion deviation from the identity function in standard error units. Two illustrations of the procedure involving SAT data were presented. Results from this kind of analysis could have implications for equating practices and test assembly practices. If the two versions of the operational SAT test had differed by more than two SEEs, it might have been necessary to compute separate equating functions for each order and to consider either taking an average of the two functions or using separate functions for each order depending on the practical consequences of each option. If the equating function for one or more of the item orders had been shown to differ by more than two SEEs, then test assemblers would have been alerted to the fact that item order affects test properties and that item order needs to be taken into account when assembling tests.

Our approach provides an objective, practical procedure for assessing test equivalence, but does not provide a formal statistical test. Statistical procedures for assessing the test score equivalence exist. Gulliksen (1950, pp. 173-190) describes a statistical test for parallel tests that requires several tests to be administered to the same group of examinees. The Kolmogorov-Smirnov two-sample test for identical score distributions has also been suggested.

References

Braun, H. I., and Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), *Test equating*. (pp. 9-49) New York: Academic Press.

Gulliksen, H. (1950). *Theory of mental tests*, New York: Wiley & Sons.

Harvey, A. L. (1987). *Effects of order of quantitative comparison items on mathematics test performance.* Manuscript in preparation.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: An historical perspective on an immediate concern. *Review of Educational Research, 55*(3), 387-413.